Information search subsystem for intellectual system of improvement of pressure processing processes

IOP Conf. Series: Materials Science and Engineering **862** (2020) 022015 doi:10.1088/1757-899X/862/2/022015

Information search subsystem for intellectual system of improvement of pressure processing processes

Sergey V Byvaltsev[1,2]

[1] Ural Federal University named after the First President of Russia B. N. Yeltsin, 19, Mira St., Ekaterinburg, 620002, Russian Federation
[2] Institute of Engineering Science, Ural Branch of the Russian Academy of Sciences, 34, Komsomolskaya St., Ekaterinburg, 620049, Russian Federation

E-mail: s_ikar@mail.ru

Abstract. The software for the information retrieval subsystem of the intelligent system for optimizing pressure processing processes was developed. The solution to the problem of finding information on the Internet comes down to finding key information. Algorithms for solving the problem and work results are presented. The advantages and disadvantages are shown in comparison with existing approaches. A program that implements a search algorithm is proposed.

1. Introduction

As it has been considered in works [1, 2] one of subsystems of intellectual system of improvement of pressure processes (further IS) is a subsystem of management of the Database and Knowledge Base. At the same time, the adequacy of IS performance of its functions depends on the hypotheses contained in it, represented by analytically mathematical models [3], recorded in general form, such as those given in the work [4], and forming the basis of the Knowledge Database. The results of modeling and full-scale experiments are automatically saved in the Database [5]. To refine the models and calculate the coefficients included in them, a huge amount of source data is required. Loading of some initial data, for example, properties of materials, represents rather monotonous routine work and often does not require creativity and high qualification of the researcher. This paper proposes to automate the process of searching information on the Internet and entering data into IS using a computer program developed by the author.

2. Task definition and solution methods

Currently, parsers are used to solve such problems, which help to find the right information in a relatively short time, in relation to the traditional search for information in the browser. The word parsing itself (from English Parsing) means syntactic data analysis, widely used in computer science [6]. The information parsing allows choosing only the necessary information on searching request, throwing out unnecessary data [7]. Such a search system is generally created as a mathematical model that compares data with formal grammar [8], [9], and is written in one of the universal programming languages [10], such as PHP, Java, Ruby, Python, Lua [11].

Today many searching programs are developed, but the sphere of their use is most often narrowly directed, as for example, the program [12].

A parser program was created for IS to search various information, ranging from numerical data to text fragments. It is universal and is developed on high level programming language Java [13] which is cross-platform and is carried out on any operating system with installed JVM (Java Virtual Machine).

Algorithm of Parser's work is as follows:

- sending an HTTP request and receiving the code of a web resource;
- reading, retrieval and processing of data received from the web resource;
- presenting the processed data in the necessary form.

Many automated search systems such as Gopher, WAIS and WWW are available for parsing information. Later on, the system was used to work with hypertext WWW. The most successful solution for the parser flexibility will be the use of open source libraries, for example, Java-libraries Jsoup [14]. This library has quite a lot of classes and methods, which allow easy enough to implement processing and extraction of information from web resources. As an integrated development environment (IDE) was used Eclipse Java program, which includes many tools for compiling and building a ready-made program [15].

The information obtained from a web resource relays to a database built using MySQL.

The program is implemented on the eighth version of Java. Synchronization with SQL database is implemented on MySQL SERVER version 8.0.15. Communication with the Internet is carried out via HTTP protocol [16]. The basic principle of information search in Jsoup library is implemented on search of key words, symbols and tags of HTML code. It also converts an HTTP request into a clear URL [17], which is obtained by conventional search using search systems such as Google, Yandex and Duckduckgo. The developed parser program searches necessary information with the help of various search systems, processes and presents to the user in a simple and convenient way. As a graphical user interface (GUI) program uses a simple one-page local site, which is created by Java program and automatically opens after the search for the necessary information. These data are recorded in the Database to simplify the search for the second request.

The concept of the developed parsing program, which is a part of information search subsystem, is shown in figure 1.

Let's consider in more details the structure of the developed program.

As it is shown on figure 2, the program consists of two main blocks.

The first main block 1 (see figure 2) is realized in Java language and consists of several Blocks.

The first Block, the structure of which is shown in figure 3, decodes all characters of the inserted request in Unicode to form an HTTP request [17]. As a result, a URL-link to search the necessary information is formed, and with the help of the above mentioned Jsoup library the request is sent to the Internet [18]. The request is sent to several search systems: Google, Yandex and Duckduckgo. The result of the query is the HTML code of the site, written in a string variable. This variable contains the necessary information that needs to be retrieved and processed.

The next step, shown in figure 4, is parsing the HTML code of the page, which has all the required information.

With the built-in features of the Jsoup library, HTML tags that contain the necessary information are found. Received and retrieved information is processed by search algorithms in the text fragments by keywords and is translated into a view understandable to the researcher.

The information obtained by MySQL is written to the IS Database (see figure 5) and is displayed in the GUI for control purposes.

The request and result are saved in the IS Database. To connect to the SQL database the built-in JDBC library was used, specially designed to connect programs written in Java to MySQL. In the process of creating a MySQL connection, the same Windows-1251 encoding is set for both Java and MySQL Server.

Further, as shown in figure 6, the formation of the graphical interface is performed.

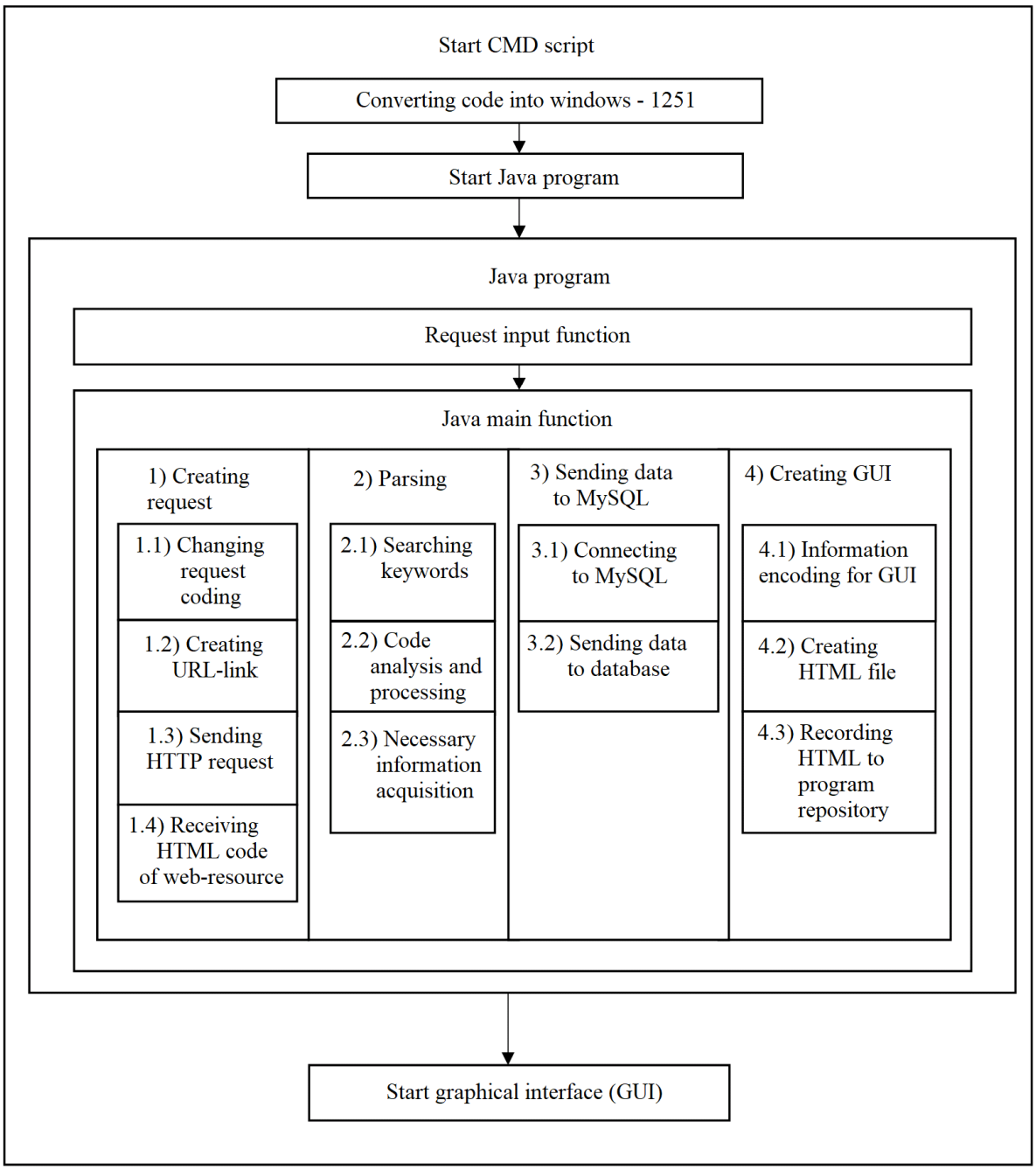

Figure 1. The concept of the information retrieval subsystem.

As a graphical interface we use a typical HTML page. The GUI is adaptable to any screen resolution, both for a personal computer and a mobile device. We have implemented such an interface using a special tool Bootstrap-4, which is used for website layout.

The program, having received the necessary data, creates an HTML page to display the results of the parser program, for which the HTML file and the necessary files are written to the root folder of the project.

To ensure cross-platform functionality of the program, the developed program is compiled and collected in a jar-archive.

The second main Unit 2 (see figure 2) is designed to provide cross-platform functionality of the developed program, as well as to start and adjust it with the necessary encoding for correct coordination of Java-program with the used operating system. For example, let's consider how to coordinate the program with the command line of Windows operating system. In this case, the program is launched using the CMD-script Windows, the code of which is shown in figure 7.

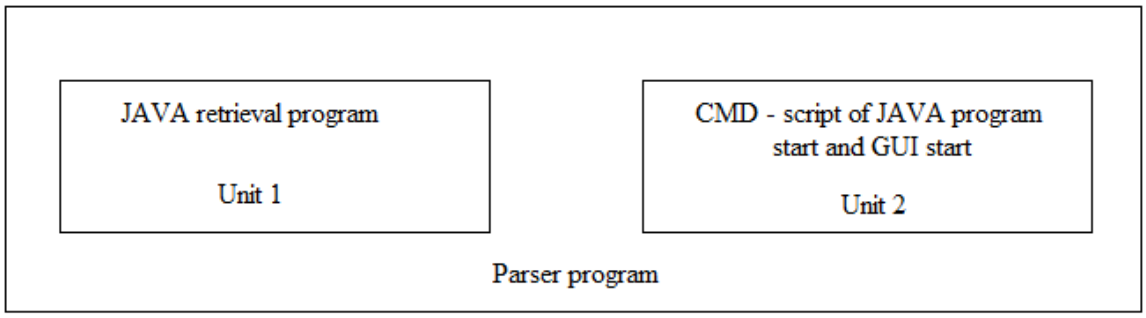

Figure 2. Parser program structure.

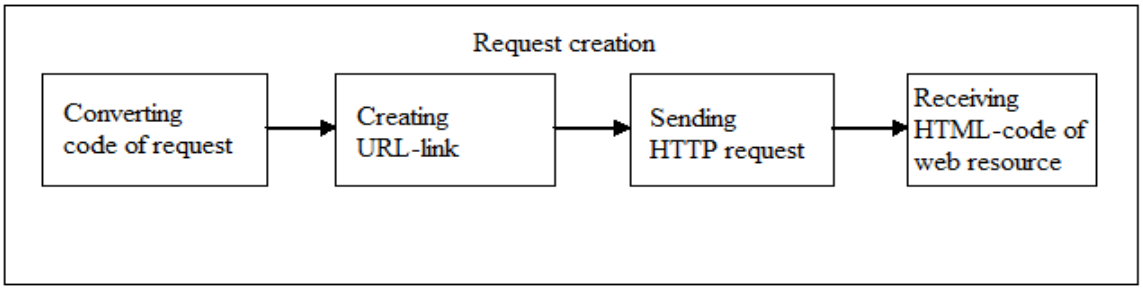

Figure 3. Structure of the request creation Unit.

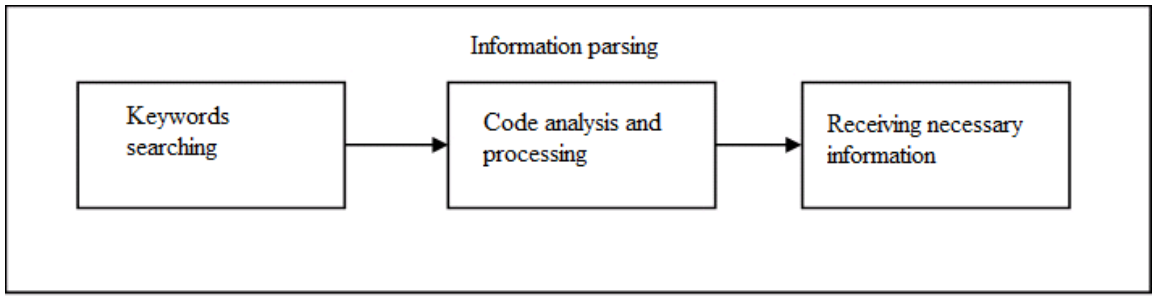

Figure 4. Information parsing.

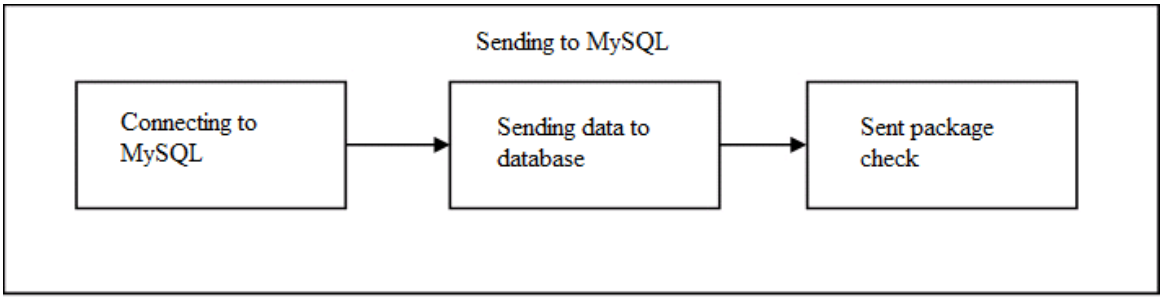

Figure 5. Sending to MySQL.

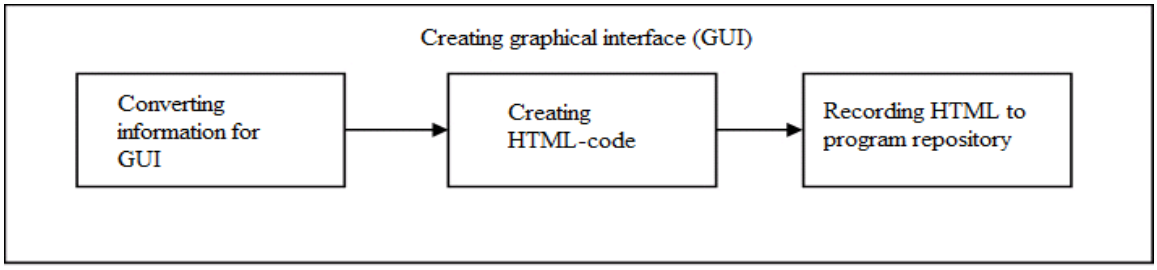

Figure 6. Creating graphical interface.

```
1 chcp 1251
2 cd C:\java\
3 java -jar java_parsing_09052019.jar
4 timeout 1
5 start C:\java\notes.html
```

Figure 7. Program launching code.

In this script for entering a request in Russian, the first line of the script changes the CMD input encoding into Windows-1251. Then, in the second line of the script code, the necessary program repository is selected. Then the JAR-file of the program is launched. After the request is entered and the program is executed, the graphical interface is launched, where the results of the parser program are given, as well as the data already written to the Database in previous requests.

As a result, the information parser program consists of three files: JAR archive, GUI (HTML page), and CMD launch script.

3. Results and discussions

To test the program, let's run the script and see that after starting the CMD-script, the necessary CMD encoding is set, the program repository is opened and the program is launched.

After entering the request, in this case it was required to find the gold melting point, the main program algorithm is launched. Each step of the algorithm is displayed on the screen to monitor the correctness of the program. After the execution is completed, a delay of one second is set so that the program can finish writing the HTML file and close the processes. The final stage is the launch of a static HTML site, which displays the current result of the program, as well as the previous search results from the SQL database. The developed program can search not only for numerical data, but also for lines of text. The result is an easy to use program that displays the necessary information in the graphical interface.

The parser is launched in the same way in other operating systems.

4. Conclusion

An automated program to collect information, which is written in the Java programming language, is developed. The main advantage of the program is the speed of information retrieval, compared to conventional search, and cross-platform system, which makes it possible to run on any computer. The program is compiled in Eclipce Java program and collected in jar-archive. Configured synchronization with a database MySQL, in which data queries and results are stored. There is a simple and easy to use local site in HTML script language, which contains the result of the program. By default, the program uses the search system Yandex and if the necessary information is not found there, the parser searches for information on search systems Google and Duckduckgo. The advantage of this program is cross-platform, using as a graphical interface HTML-site and allows you to run the program on different OS. Testing the algorithm on the example of pure metal melting from a periodic table of chemical elements

of Mendeleev, showed the following results: manual search takes 50 minutes, and with the help of the program - 9 minutes, which shows the effectiveness of the used approach.

Acknowledgements
The authors are grateful to the staff of the Department of Information Technologies and Design Automation of the Ural Federal University and the staff of the System Modeling Laboratory of the Institute of Engineering Science, Ural Branch of the Russian Academy of Sciences without whose valuable advice and critical comments it would be impossible to publish this study.

References

[1] Byvaltsev S V, Zalazinskiy A G and Shveykin V P 2019 *AIP Conf. Proc.* **2176** 030003
[2] Agapitova O Yu, Byvaltsev S V and Zalazinskiy A G 2017 *AIP Conf. Proc.* **1915** 040001
[3] Byvaltsev S V 2017 *Vestnik Uralskogo gos. univ. putey soobcheniya* **4(36)** 39-48
[4] Agapitova O Yu, Byvaltsev S V and Zalazinskiy A G 2011 *Russian J. of Non-Ferrous Metals* **52(4)** 382-7
[5] Zalazinskiy A G, Byvaltsev S V, Zalazinskaya E A and Agapitova O Yu 2010 *Elec. J. Issledovano v Rossii* **13 2010** 542-51
[6] Aho A V and Ullman J D 1972 *The Theory of Parsing, Translation, and Compiling* ed G Forsythe (NJ United States: Prentice-Hall, Inc.) **1** p 1050
[7] Medeiros S and Ierusalimschy R 2008 *Proc. of the 2008 symposium on Dynamic languages* (DLS '08) 2 1-12
[8] Redziejowski R R 2008 *Fundam. Inform.* **85** 441-51
[9] Medeiros S, Mascarenhas F and Ierusalimschy R 2011 *Brazilian Programming Languages Symposium*
[10] Medeiros S, Mascarenhas F and Ierusalimschy R 2012 Left Recursion in Parsing Expression Grammars *Science of Computer Programming* (*Preprint* http://arxiv.org/abs/1207.0443v3) **96** p 24
[11] Ierusalimschy R, de Figueiredo L H and Celes W 2006 *Lua 5.1 Reference Manual* (Rio de Janeiro: Lua.Org) p 112
[12] *Parsing sites and monitoring competitors* Available at: https://xmldatafeed.com
[13] Gosling J, Joy B, Steele G and Bracha G 2005 *The Java Language Specification 3rd Edition* (Addison-Wesley) p 684
[14] Hedley J *Jsoup Java HTML Parser, with best of DOM, CSS, and j query* Retrived from https://jsoup.org/ (accessed 03.03.2020)
[15] *Eclipse IDE for Java EE Developers* Retrived from https://www.eclipse.org/downloads/packages/release/mars/2/eclipse-ide-java-ee-developers
[16] *HTTP* Retrived from https://developer.mozilla.org/en-US/docs/Web/HTTP
[17] *URL (Java Platform SE 7)* Retrived from https://docs.oracle.com/javase/7/docs/api/java/net/URL.html (accessed 03.03.2020)